VIDEO SCENE RETRIEVAL BASED ON LOCAL REGION FEATURES

Arasanathan Anjulan and Nishan Canagarajah

Department of Electrical and Electronic Engineering, University of Bristol
Merchant Venturers Building, Woodland Road, Bristol, UK

ABSTRACT

This paper describes a novel method for content extraction and scene retrieval for video sequences based on local region descriptors. The local invariant features are obtained for all frames in a sequence and tracked throughout the shot to extract stable features. The scenes in a shot are represented by these stable features rather than features from one or more key frames. Compared to previous key frame based approaches, the proposed method is highly robust to camera and object motions and can withstand severe illumination changes. The proposed approach is applied to scene retrieval experiments and excellent performance is demonstrated.

Index Terms— scene retrieval, invariant features, video segmentation, feature extraction, local region tracks

1. INTRODUCTION

Feature extraction is an active field of research in content based video retrieval and summarization. Summarising the contents of the video is important for efficient browsing and retrieval in multimedia databases. Typically most of the existing video content extraction systems select one or more key frames as being representative of each shot; feature extraction techniques such as wavelets or Gabor filters are widely used to then extract features from these frames. However an efficient key frame selection method, which works with all kinds of videos with little redundancy, is still a difficult problem. Different imaging conditions and camera and object motions make it nearly impossible to represent a shot by a small number of frames without oversampling and thus increasing the complexity and memory requirements of the system. On the other hand, any attempt to reduce the number of key frames may result in content loss and thus a failure to properly represent the shot. Furthermore, features selected from one or more key frames are not robust enough to adequately represent the scenes in a shot.

In this work, we propose the use of local invariant region features to develop a highly accurate content extraction method for video sequences. Stable features are extracted throughout a shot rather than from a small number of key frames. We propose this approach as an alternative to the key frame method. Local regions are tracked throughout a shot with features being extracted from stable tracks. An efficient method is proposed for region tracking to avoid possible repetition of the features. The proposed framework is robust to camera and object motions and can withstand severe illumination changes, spatial editing and noise.

Early approaches in key frame selection propose typically select the first frame in each shot as the key frame [1, 2]. However one key frame per shot is not always sufficient as there can exist a number of salient changes within a shot due to camera or object motion. Conversely, Ardizzone and Cascia [3] suggest making the number of key frames proportional to the length of the shot. They propose taking a key frame for each second. This approach is likely to oversample the sequence, as the semantic content may not often change that quickly. Zhang et al [4] propose a method to extract key frames based on a similarity measure between adjacent frames. They propose selecting the first frame in a shot as the key frame and compared the following frames with the key frame for content similarity. If a significant change occurs, then that frame is also selected as an additional key frame and this process continues until the end of the shot. The idea behind this method is that any content change between frames suggests significant activity in the shot and should be represented by multiple key frames. Vermaak et al [5] suggest that key frames should be maximally distinct and individually carry the most information. Here the input video is transformed into a sequence of representative feature vectors and this representation is used to define a utility function. A key frame sequence that maximises this function is obtained by a non-iterative dynamic programming procedure.

The initial inspiration of our work is obtained from the work done by Sivic and Zisserman [6, 7]. They use local invariant region descriptors to represent key frames. They adapt text retrieval techniques for fast and efficient retrieval. Local region descriptors are vector quantized into clusters and used

The work reported in this paper has formed part of the ICBR project within the 3C Research programme of convergent technology research for digital media processing and communications whose funding and support is gratefully acknowledged. For more information please visit www.3cresearch.co.uk.

Arasanathan Anjulan and Nishan Canagarajah are with the Department of Electrical and Electronic Engineering, University of Bristol, Bristol, BS8 1UB, UK (e-mail: A.Anjulan@bristol.ac.uk, Nishan.Canagarajah@bristol.ac.uk.

as visual "words" in retrieval applications. The regions obtained in key frames are tracked and any region not lasting at least three frames are rejected. In experiments, they show good performance in scene and object matching. However their system is based on key frames and any failure in key frame extraction will affect their system. As they agree that significant change in imaging conditions may limit the performance of the system because of the limited overlapping regions among key frames. This problem however is overcome in our approach by extracting key features throughout a shot rather than extracting them only from key frames.

The rest of this paper is organized as follows. The proposed approach is described in section 2. In section 3, the experimental work carried out to demonstrate the performance of our method is presented. We conclude in section 4 with suggestions for future work.

2. PROPOSED APPROACH

To the best of our knowledge, all the existing content extraction and retrieval approaches for video sequences are based on key frames. In this work, however, stable local features, obtained throughout a shot, are used in content extraction and retrieval applications. This is because the key frame method fails when sudden changes occur in camera movement or illumination. Furthermore, features selected from one or more key frames are not robust enough to adequately represent the scenes in a shot.

Our segmentation and content extraction algorithms are based on the concept of local invariant region descriptors. A brief explantation and performance evaluation of local region extraction methods can be found in [8]. We choose *Maximally Stable Extremal Regions* (MSER) algorithm by Matas et al. [9] as it performed well with affine and illumination changes. The *Scale Invariant Feature Transform* (SIFT) [10] is used to obtain the region descriptors in our experiments, as SIFT has been proved to be robust against varying imaging conditions [11].

The MSER and SIFT algorithms are used to obtain local region features in each frame. The detected local regions within a shot are tracked based on the similarity of the region descriptors in adjacent frames. Each new track at any point within a shot is compared to the existing tracks. This enables regions to be tracked through occlusions, thus avoiding repetition of the features. Once a shot cut is detected, the stable tracked regions are summarised based on the length of the run and used as representative features for that shot. Thus in this method, a shot is represented by the stable tracked features throughout the shot rather than the features from one or more key frames. Video segmentation is an essential subpart of this approach as the feature tracking should be limited within a shot. We incorporated an approach, which was previously proposed by the authors[12], in video segmentation (Local Invariant Region Based cut detection) based on the consistency of the local regions. This approach has been proven to be more robust to camera and object motion and illumination changes compared to other video segmentation methods[12].

The extracted local regions are tracked throughout the shot based on the feature matches. Some of these regions may disappear in particular frames and then reappear later in the shot. This may happen because of occlusion or failure of the MSER algorithm due to extreme conditions. We call these tracks as discontinuous. A real example of a discontinuous track is given in Fig 1 for a shot taken from the video sequence *Tennis*. Fig 1(a) shows the starting frame and the rescaled frame part to highlight the selected region. Fig 1(b) shows the heighlighted regions in the track. The region in question is tracked from frame 585 to 588 and lost in frame 589 because of the movement of the face away from the camera. However the region reappears in frame 608 and is tracked until frame 613. Although these are two different tracks, they represent the same region, thus giving the same content information. In a content extraction system, these two tracks should be joined and considered as one track. This is achieved as follows. Each new starting track at any point in the shot is compared with all the existing tracks within that shot, to avoid possible repetition of the features due to discontinuous tracks. This also enables tracking of regions through occlusions. For example, consider a frame in the middle of a shot with n tracks, $[t_1.....t_n]$; here the length of the track t_i is m_i frames. Track t_i goes through m_i frames and each point in the track contains a 128 element SIFT descriptor vector. Therefore the i^{th} track can be summarised as, $[\mathbf{d}_1.....\mathbf{d}_{m_i}]$. where $\mathbf{d}$ represents the SIFT descriptor.

If a region descriptor, $\mathbf{d}$, obtained in the current frame does not have any matches, then it will be compared with the averaged region descriptor of all existing tracks. For the i^{th} track, the averaged descriptor, $\overline{\mathbf{d}}_i$, will be obtained as follows, $\overline{\mathbf{d}}_i = \frac{1}{m_i}\sum_{i=1}^{m_i}\mathbf{d}_i$. The non matched region vector, $\mathbf{d}$ will be compared with all existing averaged tracks to find the closest averaged track. If the distance between $\mathbf{d}$ and the closest averaged track is less than a threshold then it will be considered as a continuation of that track, otherwise a new track will be formed. In the example shown in Fig 1, the new unmatched region in frame 608 matched with the earlier track from frame 585 to 588 as shown in the figure. Therefore these two tracks are joined together and will be considered as a single track.

However if the region is not lost throughout its track, then the track is a continuous track. A real example is given in Fig 2 for a shot taken from the movie sequence *Cliff Hanger*. The region in question is continuously tracked from frame 6 to 149.

Once a shot boundary is detected, the feature vectors in the stable tracks throughout that shot will be averaged and stored. The stable tracks are selected based on the length of the tracks through frames. We select a track if it goes through at least 7 frames. If the total selected tracks is greater than 200 for any shot, first 200 most stable tracks are selected. When

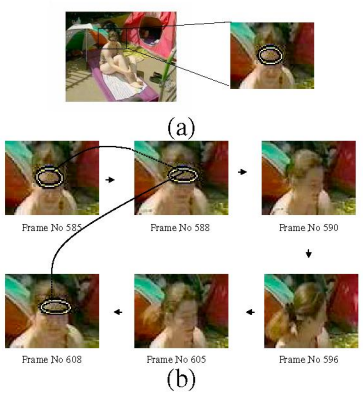


Fig. 1. An example of a discontinuous region track through occlusions (a) Starting frame of the track and the rescaled region for more clear view (b) Rescaled regions in the track. The tracked region is lost in frame 589 because of the movement of the face away from the camera. However, it reappears in frame 608 and joined with the earlier track.

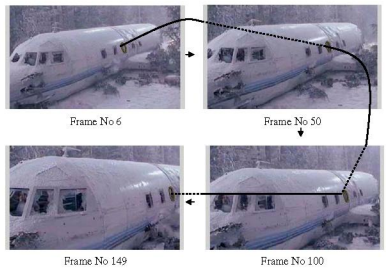


Fig. 2. An example for the continuous track for a shot taken from the movie sequence *Cliff Hanger*

a query image is presented to the system, the local region descriptors are obtained for that image and compared with the stored shot features. Based on this comparison, the best matched shots will be selected and presented as the matches. If the best match value is less than a threshold value, then no match will be possible.

3. RESULTS

The proposed approach is applied to scene matching applications. The test data uses all the shots from the movies *Run Lola Run* and *Groundhog Day*. Precision-recall curve and normalised rank measures were used in evaluating the performance.

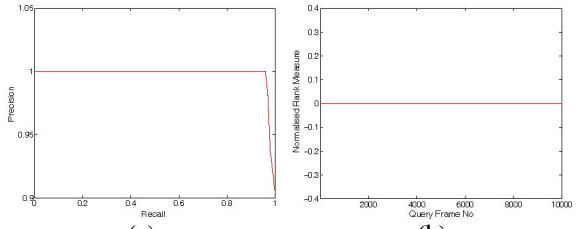
We evaluate the retrieval performance of our algorithm based on the stored stable local features. Given a query image, related shots taken of the same scene should be retrieved while avoiding other scenes. The shots may be taken under different imaging or lighting conditions, such as different camera angles, zooming positions and illumination changes.

Fig. 3. Examples of frames taken in the same scene. Each of the frame in all the sub figures is taken from different shot taken in the same scene. The frames vary significantly both in terms of the imaging conditions and the areas covered.

Furthermore a shot may cover a large area varying from one place to another and the system should be able to handle these variations. Scenes appearing in movies *Run Lola Run* and *Groundhog Day* are used in scene retrieval experiments which is often used by other researchers. In these movies, the same scenes were filmed a number of times in different imaging conditions, making these ideal video sequences for scene retrieval experiments. The ground truth of the similar scenes are selected manually throughout the whole movies. If a similar place (for example building or road) appears in different shots, we conclude them as similar shots. Examples of frames from similar shots are given in Fig 3 (a)-(d). Each sub figure contains frames taken from similar shots. As seen in the figure, the frames vary significantly both in terms of the imaging conditions and the areas covered.

A frame which contains the scene in question is given as the query. The SIFT features are extracted from the selected MSER regions throughout the frame and compared with the features from all the shots. For normalisation the total number of matched features are divided by the number of features obtained from the query frame. If the normalised value is greater than a threshold, it is presented as one of the matched shots and all the matched shots are ordered in the descending order of normalised matched value.

We use the average PR curve and average normalised rank measure [6] to evaluate the performance of our approach. Fig 4(a) shows the average PR curve. We randomly selected 10000 frames (from movies *Groundhog Day* and *Run Lola Run*) as the query image for the system and the matched shots are ob-



(a) (b)

Fig. 4. (a) Average Precision-Recall curve obtained in scene matching applications. 10000 randomly selected frames were used in the experiment. (b) Normalised rank value is plotted for all 10000 randomly selected frames used in precision recall experiment. The rank value is 0 for all the 10000 frames which indicates that all the relevant shots are retrieved as first matches for all the query frames.

tained. The precision value is calculated as the ratio of the number of correctly retrieved shots to the total number retrieved shots; the recall value is calculated as the ratio of the number of correctly retrieved shots to the number of relevant shots in the database. It is important to note that the scene in some of the selected query frames may appear only in one shot. As seen in the Fig 4(a), our algorithm gives a nearly perfect performance (precision value is more than 0.90 for any recall value). Given an image as query, our algorithm picks up all the shots in the same scene while avoiding any false alarms.

The average normalised rank of the relevant shots can be defined as follows,

$$\widetilde{Rank} = \frac{1}{NN_{rel}}\left(\sum_{i=1}^{N_{rel}} R_i - \frac{N_{rel}(N_{rel}+1)}{2}\right) \qquad (1)$$

where N is the number of total shots, N_{rel} is the number of relevant shots and R_i is the rank of the relevant shot. $\widetilde{Rank}$ is zero if all relevant shots are returned first. The $\widetilde{Rank}$ measure varies between the range 0 and 1 with 0.5 corresponding to random retrieval.

The rank measure is plotted for all 10000 randomly selected frames used in the PR curve experiment, in Fig 4(b). As clearly seen in the figure, the rank value is 0 for all these frames. This indicates that all the relevant shots are retrieved as first matches for all the query frames.

4. CONCLUSION

A novel content extraction method for video sequences based on local region descriptors is proposed. Unlike the previous approaches, the proposed method does not depend on key frame selection. Local regions are obtained and tracked throughout a shot and the stable features are used to form shot representation. The above shot representation gives better results compared to the conventional key frame method and excellent performance is shown in scene matching applications. The proposed method is robust to camera and object motions and can withstand severe illumination changes. The performance is evaluated by scene matching experiments with the movies *Run Lola Run* and *Groundhog Day*. Future work will consider identifying individual objects in video sequences based on local region descriptors and the current framework provides the foundation for this extension.

5. REFERENCES

[1] Herng-Yow Chen and Ja-Ling Wu, "A multi-layer video browsing system," *IEEE Trans on Consumer Electronics*, vol. 41, pp. 842–850, 1995.

[2] B. Gunsel and A.M. Tekalp, "Content-based video abstraction," *In Proc. ICIP*, pp. 128–132, 1998.

[3] E. Ardizzone and M.L. Cascia, "Video indexing using optical flow field," *In Proc. ICIP*, pp. 831–834, 1996.

[4] H.J Zhang, Zhong, and S.W Smoliar, "An integrated system for content-based video retrieval and browsing," *Pattern Recognition*, vol. 30, pp. 643–658, 1997.

[5] J. Vermaak, P. Peraz, M. Gangnet, and A. Blake, "Rapid summarisation and browsing of video sequences," *In Proc. BMVC*, pp. 424–433, 2002.

[6] Josef Sivic and Andrew Zisserman, "Video google: A text retrieval approach to object matching in videos," *In Proc. ICCV*, 2003.

[7] Josef Sivic and Frederik Schaffalitzky Andrew Zisserman, "Efficient object retrieval from videos," *In Proc. EUSIPCO*, 2004.

[8] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *Technical report, University of Oxford*, 2004.

[9] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," *In Proc. BMVC*, p. 2002, 384-393.

[10] D.G. Lowe, "Distinctive image features from scale-invariant key points," *Int. Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[11] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *In Proc. CVPR*, pp. 257–263, 2003.

[12] Arasanathan Anjulan and Nishan Canagarajah, "Invariant region descriptors for robust shot segmentation," *Accepted for the Proc. of IS&T/SPIE, 18th Annual Symposium on Electronic Imaging*, 2006.